Logistic regression via MCMC algorithm

Data

The data is available from Brittany Parker PSYC 497 Spring 2009. An observational study was conducted of seat belt use in Horry County, South Carolina. Vehicles were observed passing on country roads and urban streets. Its type, gender of driver, and whether driver was wearing a seat belt were recorded. The data frame contains all 400 individual cases. The response variable 'seat belt use' is binary, and all three predictor variables are categorical.

Lets look at the data structure:

```
##  Seatbelt    Vehicle     Location      Gender
##  no :126   car   :151   rural:200   female:156
##  yes:274   pickup:108   urban:200   male  :244
##            SUV   :101
##            van   : 40
```

Based on the output, it is reasonable to say that

- 'Vehicle' may not be too useful due to poor variability;

- 'Gender' and 'Location' reasonably well balanced.

To understand the data structure more precisely, the pie charts of proportion of drivers who were and were not wearing a seat belt are given in Figure 1.

Methods. Models

Inference for Proportions

The proportions from two independent binomial distributions could be compared. Lets consider a proportion of drivers who were not wearing a seat belt in urban location and that in rural location. After assigning Beta priors to each proportion and using initial beliefs with BetaBuster, the 'matching' Beta priors could be found.

It is sounds reasonable that prior mode for proportion of drivers who were not wearing a seat belt in urban location is 0.2, and suppose that it is smaller than 0.5 with 95% confidence. Also assume that prior mode for proportion of drivers who were not wearing a seat belt in rural location is 0.35, and it is smaller than 0.6 with 95% confidence. Using BetaBuster, the 'matching' Beta priors were found and the following Openbugs model was fitted:

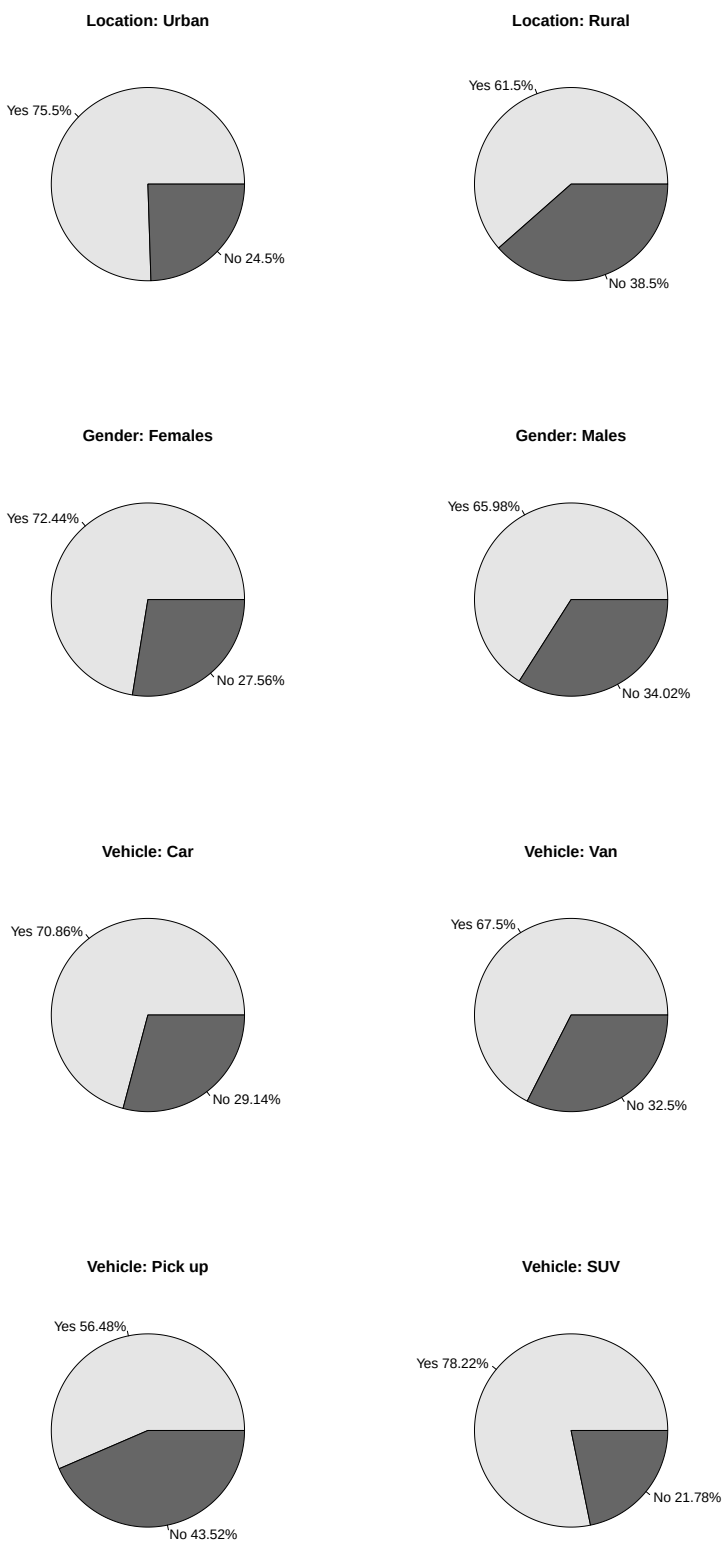

Figure 1: Pie charts

```
model{
 y1  ~ dbin(theta1,n)
 y2  ~ dbin(theta2,n)
 theta1 ~ dbeta(2.637, 7.548)
 theta2 ~ dbeta(4.759, 7.981)
 theta.ratio <- theta2/theta1
 theta.diff <- theta2 - theta1
 prob1 <- step(0.3 - theta1)
 prob2 <- step(0.45 - theta2)
}
```

As shown in the Figure 2, both informative priors used have no impact on the posterior distribution.

The Openbugs model will be run, and the Bayesian inferences about relative risk ratio, risk difference, and posterior probabilities will be given in the next section.

Logistic regression via MCMC algorithm

A binomial logistic regression was ran for these data starting with a full model:

$$log\left(\frac{Seatbelt}{1-Seatbelt}\right) = \beta_0 + \beta_1 Vehicle + \beta_2 Location + \beta_3 Gender$$

Based on the glm() output, insignificant variables 'Vehicle', and then 'Gender' were excluded from the model. The output below shows that 'Location' is very significant, and this model could be considered:

$$log\left(\frac{Seatbelt}{1-Seatbelt}\right) = \beta_0 + \beta_1 Location$$

A classical regression approach leads to the following final model:

$$log\left(\frac{Seatbelt}{1-Seatbelt}\right) = 0.4684 + 0.6571 \times Location$$

```
##
## Call:
## glm(formula = Seatbelt ~ Location, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6772  -1.3817   0.7497   0.9860   0.9860
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

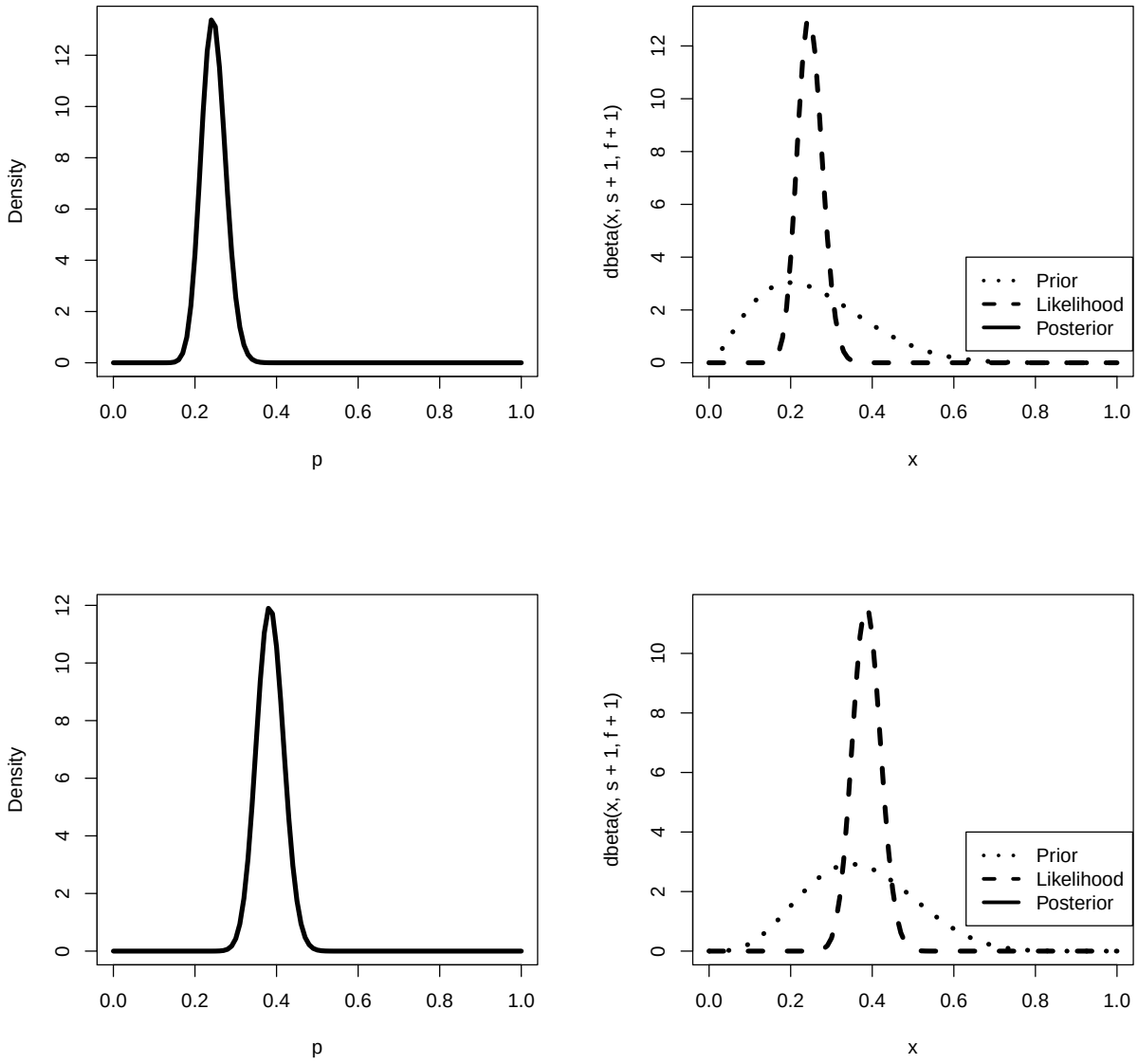

Figure 2: Prior, Likelihood, and Posterior densities

```
## (Intercept)     0.4684     0.1453   3.223  0.00127 **
## Locationurban   0.6571     0.2194   2.995  0.00275 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.43  on 399  degrees of freedom
## Residual deviance: 489.29  on 398  degrees of freedom
## AIC: 493.29
##
## Number of Fisher Scoring iterations: 4
```

Since a classical regression approach makes sense, using Rjags and R language, a binomial logistic regression model of seat belt use in Horry County, SC will be applied. The following four binomial logistic regression models will be applied to the data and compared using Bayes Factors and Deviance information criterion.

Model 1:

```
model.full.1 <- "model{
  for(i in 1:n) {
    logit(p[i]) <- b0 + b1*Vehicle[i] + b2*Location[i] + b3*Gender[i]
    y[i] ~ dbin(p[i], 1)
}
b0 ~ dnorm(0, 0.0001)
b1 ~ dnorm(0, 0.0001)
b2 ~ dnorm(0, 0.0001)
b3 ~ dnorm(0, 0.0001)
}"
```

Model 2:

```
model.reduced.2 <- "model{
  for (i in 1:n) {
    logit(p[i]) <- b0 + b1*Vehicle[i] + b2*Location[i]
    y[i] ~ dbin(p[i], 1)
  }
        b0 ~ dnorm(0, 0.0001)
        b1 ~ dnorm(0, 0.0001)
        b2 ~ dnorm(0, 0.0001)
}"
```

Model 3:

```
model.reduced.3 <- "model{
  for (i in 1:n) {
    logit(p[i]) <- b0 + b1*Gender[i] + b2*Location[i]
    y[i] ~ dbin(p[i], 1)
  }
        b0 ~ dnorm(0, 0.0001)
        b1 ~ dnorm(0, 0.0001)
        b2 ~ dnorm(0, 0.0001)
}"
```

Model 4:

```
model.reduced.4 <- "model{
  for (i in 1:n) {
    logit(p[i]) <- b0 + b1*Location[i]
    y[i] ~ dbin(p[i], 1)
  }
        b0 ~ dnorm(0, 0.0001)
        b1 ~ dnorm(0, 0.0001)
}"
```

Analysis. Result

MCMC for the full model 1 was ran. The two distributions ('n.chains=2') were sampled. 'n.burnin' was set to 5000 to discard a sizeable number of early values to help guarantee a better posterior. To increase sampling efficiency and avoid high autocorrelation 'n.thin' was increased here. Keeping out of every $k = 10$ samples for our distribution helps effect randomness. Finally, 'n.iter' was specified to 30000, this amount of values was kept for the posterior distribution, after thinning and discarding burn-in values. The MCMC output is given below. Gelman-Rubin measures whether there is a significant difference between the variance within and between several chains. The *gelman.diag* gives the scale reduction factors for each parameter of 1 which means that between chain variance and within chain variance are equal, which suggests adequate convergence.

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 400
##    Unobserved stochastic nodes: 4
##    Total graph size: 2453
##
## Initializing model
##           Mean        SD         50%       2.5%      97.5%
## b0  0.66037842 0.2250518  0.65893152  0.2252757 1.10715322
## b1  0.03939513 0.1116959  0.03828284 -0.1772884 0.26022516
```

```
## b2  0.69704944 0.2255043  0.69575510  0.2579358 1.13831701
## b3 -0.38945150 0.2317954 -0.38792062 -0.8455787 0.05438499
## Potential scale reduction factors:
##
##    Point est. Upper C.I.
## b0          1          1
## b1          1          1
## b2          1          1
## b3          1          1
##
## Multivariate psrf
##
## 1
```

Based on the Figures 3 and 4, the trace and density plots indicate no concern with convergence. A visual inspection of the chains suggests they are mixing well. The high autocorrelation issue is still presented here. The possible explanation of it is that the likelihood becomes messy after fitting log model, and MCMC shoud be run much longer.

Based on the output, the 95% credible intervals corresponding to the β_1 and β_3 regression coefficients include zero, which suggests they do not explain the variability of the seat belt usage.

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 2
##    Unobserved stochastic nodes: 2
##    Total graph size: 19
##
## Initializing model
##                   Mean         SD       2.5%     97.5%
## prob1       0.9583333 0.19984297 0.00000000 1.0000000
## prob2       0.9691667 0.17288036 0.00000000 1.0000000
## theta.diff  0.1384592 0.04498528 0.05020206 0.2266774
## theta.ratio 1.5867132 0.24287343 1.17563373 2.1248453
## theta1      0.2458417 0.02986034 0.18907860 0.3077382
## theta2      0.3843009 0.03400471 0.31912089 0.4524414
```

Based on the ouput above, the estimated posterior mean for the proportion of drivers who were not wearing a seat belt in urban location ($\hat{\theta_1}$) is 0.246, and the estimated posterior mean for that in rural location ($\hat{\theta_2}$) is 0.384. Based on the risk difference and relative risk ratio values, θ_2 is greater than θ_1 on 0.138, and θ_2 is 1.5 times larger that θ_1. Also there is a 95.8% chance that proportion of drivers who were not wearing a seat belt in urban location is less than 0.3, and there is a 96.9% chance that proportion of drivers who were not wearing a seat belt in rural location is less than 0.45.

The deviance information criterion and Bayes factors are used to compare the fit of the four models. Like AIC and BIC, lower DIC values indicate better fit. The plausibility of

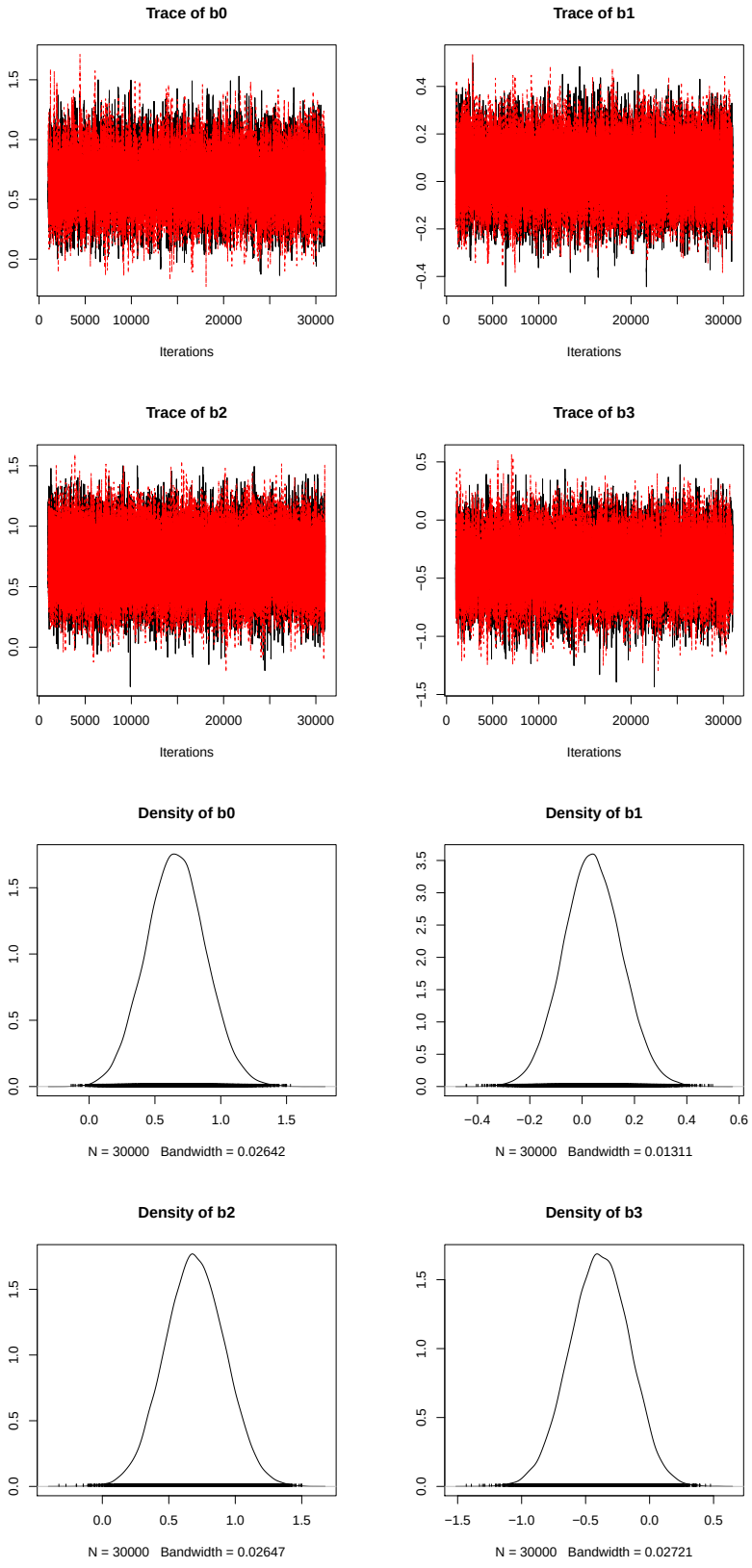

Figure 3: Trace and density plots for model 1

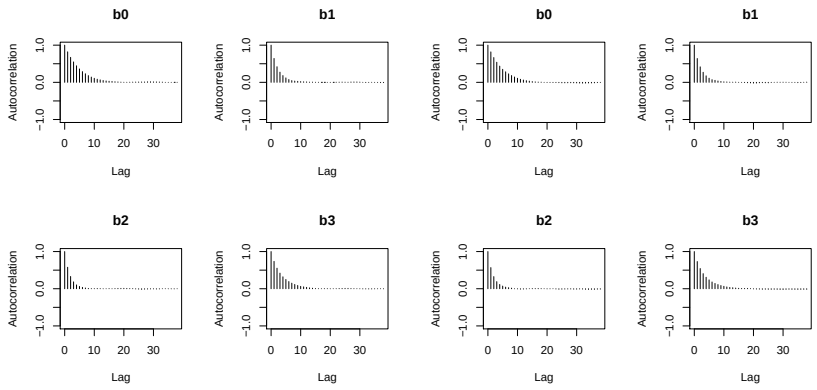

Figure 4: Autocorrelation plot for model 1

the two different models M_1 and M_2 is assessed by the Bayes factor, which is equal to the ratio of posterior probabilities of the two competitive models:

$$B = \frac{P(D|M_1)}{P(D|M_2)}$$

.

Model	DIC	BF: model.4/model.n
1. logit(Seatbelt) ~ Vehicle + Location + Gender	490.9	8.2
2. logit(Seatbelt) ~ Vehicle + Location	492.1	6.1
3. logit(Seatbelt) ~ Gender + Location	488.9	1.7
4. logit(Seatbelt) ~ Location	491.2	1

Based on the DIC, the model (3) should be choosen; BF indicates that model (4) is the most appropriate model.Thus, the reduced model (4) could be fitted.

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 400
##    Unobserved stochastic nodes: 2
##    Total graph size: 1613
##
## Initializing model
##         Mean        SD     Naive SE Time-series SE      2.5%     97.5%
## b0 0.4715149 0.1454119 0.0005936418    0.001205385 0.1874157 0.7601359
## b1 0.6601200 0.2199296 0.0008978590    0.001812346 0.2289109 1.0942634
## Potential scale reduction factors:
##
##    Point est. Upper C.I.
## b0          1          1
## b1          1          1
##
```

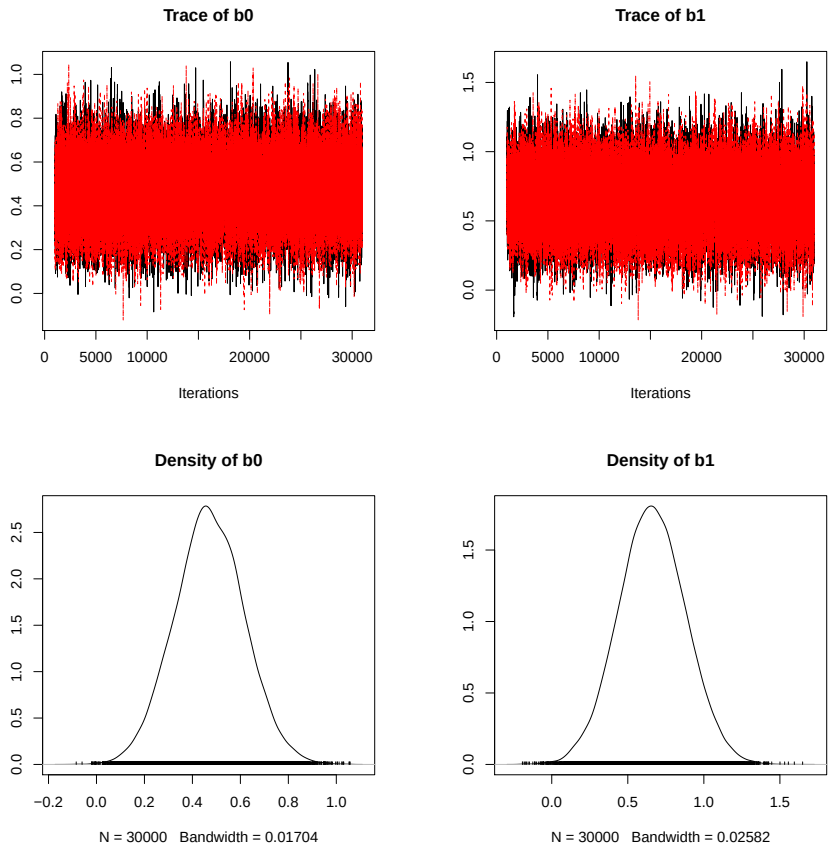

Figure 5: Trace and density plots of simulated draws of parameters for chains for model 4

```
## Multivariate psrf
##
## 1
```

Figures 5 and 6 show the diagnostic plots for model 4. The trace and density plots look good. The autocorrelations are high for the first lags but reduce slowly as a function of the lag. Based on the output, the 95% credible intervals corresponding to the β_0 and β_1 regression coefficients don't include zero, which suggests they are contributing to variability of the seat belt usage. Thus, the model (4) is our final model:

$$log\left(\frac{Seatbelt}{1-Seatbelt}\right) = 0.472 + 0.66 \times Location$$

The logistic regression equation yields the model expressed in terms of the odds:

$$\frac{P(Seatbelt = 1)}{1 - P(Seatbelt = 1)} = e^{0.472}(e^{0.66})^{Location}$$

That is, in urban location the estimated odds of people wearing a seat belt is roughly $e^{0.472+0.66} = 3.1$. The estimated odds of people wearing a seat belt in rural location is approximately $e^{0.472} = 1.6$.

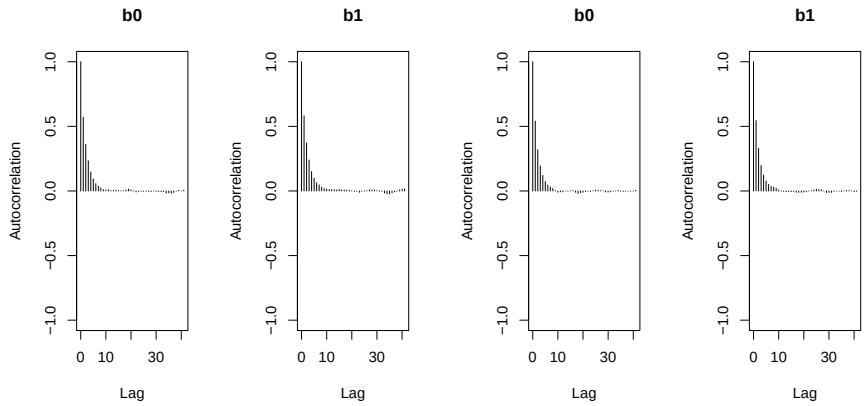

Figure 6: Autocorrelation plot of simulated draws of parameters for each chain

References

1. Albert, J., (2009) Bayesian computation with R. Second Edition. Springer Verlag (A book). May download this book through SpringerLink via University Libraries.

2. Christensen, R., Johnson, W., Branscum, A. and Hanson, T.E. (2010) Bayesian Ideas and Data Analysis. An Introduction for Scientists and Statisticians. Chapman and Hall/CRC Press.

3. Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003), Bayesian Data Analysis, New York: Chapman and Hall.

4. Hilbe, J., (2015) Practical Guide to Logistic Regression. Chapman and Hall/CRC Press.

5. Kruschke, J., (2010) Doing Bayesian Data Analysis: A Tutorial with R and BUGS. Academic Press/Elsevier.